A HIERARCHICAL CLUSTERING TECHNIQUE

Earnest Wells Richardson

United States
Naval Postgraduate School

THESIS

A HIERARCHICAL CLUSTERING TECHNIQUE

by

Earnest Wells Richardson

Thesis Advisor: J. R. Borsting

September 1971

Approved for public release; distribution unlimited.

A Hierarchical Clustering Technique

by

Earnest Wells Richardson
Lieutenant Commander, United States Navy
B.S., State College of Arkansas, 1962

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1971

ABSTRACT

This thesis summarizes statistical clustering procedures and presents in some detail a hierarchical clustering technique and computer routine which utilizes Euclidean distances as measures of object similarity. An application of the technique is made to scores derived from a qualitative data base describing mentally disturbed children, and results of the application are compared to results obtained from previous clustering studies.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

A. PURPOSE AND SCOPE

The primary purpose of this paper is to apply a hierarchical clustering technique to qualitative data. Statistical clustering procedures are discussed in general and a computer-based hierarchical clustering scheme is presented. The technique is applied to data provided by Dr. Bernard Rimland of the U. S. Naval Personnel Research Laboratory and Institute for Child Behavior Research, San Diego, California, and comparisons of the results of this application to results previously obtained by other clustering techniques are summarized.

B. DISCUSSION OF CLUSTERING PROCEDURES

Whereas discrimination is the broad term used to describe the problem of assignment of the objects under study to one of several known groups based on comparisons of known characteristics of the groups and the measured characteristics of the objects, clustering techniques employ only a priori selection of measures of similarity and are designed to find an inherent structure solely from the data. In general, specific characteristics of the different groups are not available and neither the number of different groups nor their relative frequency of occurrance is known.

These techniques can be applicable for any situation in which it is desired to discriminate between groups of objects and when the researcher is not willing to assume that his

knowledge of class membership is sufficient to guide the grouping procedure, or when it is desired to explore the underlying structure of objects solely on the basis of interobject similarity. Ball [Ref. 1] provides references to articles describing applications of clustering techniques to such diverse disciplines as: archeology, geography, economics, electrical engineering, information retrieval, market analysis, medicine and psychology. Clustering also has military application in such fields as personnel classification, system evaluation, and pattern recognition. In detection systems, for instance, the detection characteristics can be used as the measurements for a clustering technique. to determine how well these characteristics provide "natural" discrimination between targets and other contacts.

A brief discussion of research on autistic children is included to provide background information on the data used in applying the clustering procedure. A qualitative data base complicates the clustering problem in that "similarity" between two different responses can only be determined by very close examination of the states of the measurements. The problems presented by the qualitative nature of the data have been avoided by utilizing various derivatives of the data as the measurements of interest.

Clustering has been defined simply by Ball [Ref. 1] as "the finding of data-derived groups on the basis of the groups being internally similar." Other terms used to describe these procedures include clumping, partitioning, and decomposition

of mixtures. The term "numerical taxonomy" normally applies to computer-based techniques and has been used primarily in conjunction with biological studies. With the increased availability of high-speed electronic computers, researchers in a variety of disciplines have recently developed or utilized classification procedures; but because the variety of application is great and the literature scattered, it is difficult to know what techniques exist. Ball [Ref. 1] provides a summary of many techniques, offers a framework within which the methods can be organized and includes an extensive bibliography for clustering and discrimination.

Solomon [Ref. 2] lists three major avenues of approach in solution to a clustering problem:

1. Total enumeration of all data partitions and the subsequent selection of a good or optimal clustering configuration.
2. A stepwise clustering scheme that selects for each number of clusters the best available groupings with the realization that it may ignore some good configurations in the process.
3. Reduction of multivariate data to two or three orthogonal dimensions, producing a graphic or pictorial representation that permits visual clustering.

An essential step in any of these approaches is representation of the data and establishment of measures of similarity. Since the choice of the variables to be studied, their interrelationships and the measures of similarity are the basis

for any clustering scheme, much consideration must be given to ensure that "closeness" in the sense of the similarity measures indicates closeness in the sense of the objectives of the study. The simplest and most common measures of similarity are those which combine the effects of individual variables into a single number. This assumption of numerical comparability allows clustering processes that group objects by overall similarity. Ball [Ref. 1] lists five types of similarity measures:

1. Association: The similarity between object X and object Y is the number or a function of the number of variables for which X and Y have the same response.
2. Correlation: Correlation between object X and object Y is a function of the angle between their respective vectors. It is most useful when the pattern of ratios of the variables is the prime determinant of similarity.
3. Distance: Many different distance measures are available. Weightings can be applied to absolute or Euclidean distances and can be derived either from an a priori evaluation of each variable's importance or from the data, as in Mahalanobis weightings. Several distance functions are discussed by Ball.
4. Probabilistic: These measures are used primarily when it is appropriate to modify weights of the variables on the basis of population statistics.

5. Functional: For functional measures, the value of similarity is a function of the distance from other objects.

When measures of similarity between objects have been established, the measures must be modified to provide meaningful similarity between groups of objects and between objects and groups.

The first approach (evaluation of all possible configurations) will obviously yield the "best" grouping. However, even with the present state of computer technology, this type of procedure is usually infeasible. Fortier and Solomon [Ref. 3] point out that there are 1,709,751,003,480 distinct partitions of 19 objects into eight clusters. To evaluate all partitions for 1, 2, ..., 19 clusters is inconceivable in almost any situation. In the same paper, results of their attempts at random sampling of the distinct partitions are discussed. These experiments were disappointing and it is pointed out that in most clustering situations, there are many "poor" or "not good" solutions and a minute number of good solutions. Unrestricted random sampling does not appear to provide a reliable means of avoiding the total enumeration process.

The third approach offered by Solomon (reduction of the data) is essentially a statistical procedure to be performed prior to the application of clustering techniques. The dimensionality of the measurement vectors may be reduced through factor analysis or principal components and clustering

techniques applied using these factor scores or components as the variables of consideration. In situations when the measurement vector is large, these procedures are justified. However, interpretation of the meaning of these scores is often difficult and clustering based on these variables may leave the researcher with a problem of determining what characteristics cause the resulting cluster configurations.

One of the most practical approaches in arriving at natural data derived configuration applies stepwise or iterative schemes; or procedures which eliminate most of the poor solutions before the clustering process begins. Fortier and Solomon [Ref. 3] propose a technique which eliminates in advance most of the poor solutions encountered in the total enumeration process. With this procedure, some prior knowledge of object similarity must be available, since the assumption is made that two objects should be in the same cluster if their similarity measure is greater than a preassigned constant. In the contrary case, they should not be in the same cluster. The PROMENADE system [Ref. 4] uses an interactive computer system with a graphic display, which allows the researcher to interactively control the clustering algorithms and eliminates investigation of some of the poor partitions.

Other specific clustering techniques have been developed which avoid the total enumeration process. These normally use some form of the iterative approach, whereby initial cluster points are established and with each iteration, some

reorganization of clusters and objects is accomplished based on existing object and cluster similarity. The clustering technique discussed in the following section utilizes these stepwise procedures.

II. A HIERARCHICAL CLUSTERING SCHEME

A. DISCUSSION

The clustering technique described in this section is a form of the class of clustering procedures termed hierarchical schemes by Johnson and Ward [Refs. 5, 6]. Ball [Ref. 1] refers to this type of procedure as clumping and points out its analogy to nearest neighbor methods. A similarity (or distance) matrix based on measurements in Euclidean space is initially established depicting similarity between all pairs of objects to be clustered. On the first iteration, the two most similar objects are combined to form the first cluster, an average Euclidean point is computed, and a similarity matrix between the unclustered objects and the cluster is established. On subsequent iterations, comparisons are made to determine the pair of items (two objects, two clusters, or an object and a cluster) which are most similar. This selection leads to combining two objects to form a new cluster, combining two clusters to reduce the number of existing clusters, or adding an unclustered object to one of the existing clusters. The necessary average positions are recomputed and similarity matrices between clusters and between unclustered objects and existing clusters are updated. The process is repeated until all objects are placed into a single cluster.

Each of the unclustered objects can be considered as a cluster containing one object and if there are n total objects to be clustered, the result of the scheme is to provide clustering arrangements for n, n-1, n-2, ..., 1 clusters. In some clustering situations, it is conceivable that the researcher has some notion, based on previous studies or experience in the field, as to how many separate groups should be formed. In other situations, it may be possible to develop mathematical functions of the similarity of object groupings to be optimized. This procedure is discussed in some detail by Ward [Ref. 6]. In other situations, the researcher may still be faced with the problem of determining which of the arrangements provides the "best" natural partitioning of the objects; he must view the arrangements as simply a tool by which to examine the characteristics inherent in the different clusters. For any iterations, the combining of two objects, clusters, or an object and a cluster is a result of their similarity and this similarity should provide some measure of worth for that iteration. If for four subsequent iterations, the similarities used are .085, .016, .136, and 1.27, some indication is given that the fourth iteration combined two relatively dissimilar items and the clustering configurations near this iteration should be investigated. These numbers are the similarities from the validation problem used to test the stepwise procedure and the computer routine.

An English language flow chart depicting the basic logic of the computer program and a copy of the program statements

are included as Appendices A and B, respectively. The computer routine can be used for various clustering situations in which the measurements can be considered as, or can be scaled so that, Euclidean distance measures yield meaningful measures of object similarity, by modifying six of the program statements: The DIMENSION statement must reserve ample storage space for required vectors and arrays; DIMENSION D(I,I), XND(I,K), NCL(J,I), DB(J,J), DC(I,K), XSM(J), XSU(J), XST(K,J)

where: I = Total number of objects to be clustered.

J = Measurement space on which clustering is to be based.

K = Maximum number of clusters needed.

Establishment of the parameter values affecting the size of the program is accomplished by setting:

TT = Number of objects.

CC = Number of clusters needed.

SS = Measurement space.

The FORMAT statements for reading the measurement data and for the printout require modification to conform to data input format and the number of objects in the study. In situations where the measurement space is large or the individual measurements are represented by large numbers, the scores may require scaling so that the similarity between objects does not exceed 88880.

B. VALIDATION

Data from Fisher's [Ref. 7] classic Iris problem was utilized in an attempt to validate the clustering technique and computer routine. Measurements of four characteristics of 50 flowers, each of three species of Iris plants, were used as the discriminators and to describe the axes in Euclidean four-space. Means for the four measurements are given in the following table:

	Iris Setosa	Iris Versicolor	Iris Virginica
Sepal length	5.006	5.936	6.588
Sepal width	5.428	2.770	2.974
Petal length	1.462	4.260	5.552
Petal width	0.246	1.326	2.026

TABLE I: MEASUREMENT MEANS OF IRIS DATA

The Setosa and Versicolor varieties were found growing together in the same location, but the sample of the third species (Virginica) differs in that it was taken from a different natural colony--a circumstance which might considerably disturb the mean values. Fisher reported that works of botanists of the period suggested the interesting possibility that Versicolor was actually a hybrid of the other two species and suggested that if this were true, the Virginica exerted a slightly preponderant influence.

At a point in the scheme indicated by the iteration similarity procedure, the technique correctly identified the data as coming from three populations. All 50 of the

Setosa variety were placed into one cluster and none of either of the other varieties was present. A second cluster contained 36 plants (all Virginica) and the third cluster contained 50 Versicolor and the remaining 14 of the Virginica species. The 14 "misclustered" Virginica plants have measurement means of 6.05, 2.71, 4.94, and 1.84; two of which are actually closer to the Versicolor mean. Because of the similar measurement scores of the Versicolor and Virginica species, it is felt that the results of the test problem provide some verification of the worth of the step-wise procedure.

III. DEVELOPMENTS IN AUTISTIC CHILDREN RESEARCH

A. INTRODUCTION

In order to examine the feasibility of using an iterative approach to clustering and to hopefully provide some useful results for a deserving problem, an attempt is made to cluster mentally disturbed children. This section provides background information on the problems associated with classification of these children and describes the available data set.

B. HISTORY

In 1943, Dr. Leo Kanner, then Director of the Child Psychiatry Clinic at Johns Hopkins Hospital, published his first paper on emotional disorders in atypical children [Ref. 8]. A year later in another paper, he named the new syndrome Early Infantile Autism [Ref. 9]. In these papers, he reported the presence of this disturbance in behavior in early infancy; a strange but common pattern of motor and language behaviors, behavior of both genius and idiocy, and complete absence of any evidence of physical or neurological defect.

Kanner and other forerunners in the field of child psychiatry started research on children exhibiting these symptoms. Despite the voluminous literature that has been written on the subject, the origin of the disease and reliable cures are still a mystery. Investigations have unveiled a

large variability in the symptoms displayed by children who are classified as autistic and because of the lack of objective diagnostic methods and various interpretations of the diagnostic terms used, this variability has led each researcher to notice different significant traits in the children.

Rimland [Ref. 10] defines autistic children as good-looking, highly skillful, intelligent-appearing young children who resist change and treat people as if they were objects. Bender [Ref. 11] sees autism as a defense mechanism to avoid dealing with the world's demands, and therefore as a symptom shared by the retarded as well as the highly sensitive young child. Other psychiatrists such as Bettelheim and O'Gorman [Refs. 12, 13] stress other sets of symptoms: child's retreat into isolation as a reaction to parental cruelty or indifference and the existence of multiple etiologies being a particular symptom syndrome, respectively.

In 1961, a committee of British psychiatrists headed by Creak [Ref. 14] devised a general criteria list for diagnosing a child. To be labeled psychotic, the child had to display a number of the following symptoms:

1. Impairment of emotional relationships with people.
2. Apparent unawareness of the child's own personal identity.
3. Pathological preoccupation with certain objects.
4. Sustained resistance to change in the environment.
5. Abnormal perceptual experiences.
6. Anxiety.

7. Speech abnormalities, in particular speech not used for communicative purposes.
8. Disturbances in mobility patterns.
9. Background of serious retardation on which special skills or islands of intelligence are superimposed.

It was not expected that a child show all nine symptoms.

To avoid the unclear diagnostic terminology and conflicting etiological opinions, a significant data base on a large sample of emotionally disturbed children was needed. They could be categorized into homogeneous groups with respect to symptom syndromes and then each syndrome could be related to the suspected etiological variables. Researchers and scientists are currently seeking the causes and cures of childhood psychosis, but are still greatly limited by the lack of objective diagnostic methods. Modern advances in computer technology have provided the opportunity for progress toward meaningful ways of dividing children with emotional disorders into homogeneous groups, so that meaningful scientific work can be done. By applying pattern analysis techniques to large amounts of data on psychotic children, it is hoped that it will be possible to find groups of children who exhibit very similar behavioral characteristics. It is expected that further research on the homogeneously grouped children will aid in the solution to problems of cause and cure.

C. CURRENT DATA BASE

Dr. Rimland's Diagnostic Checklist for Behavior-disturbed Children (included as Appendix C) first appeared in his book Infantile Autism [Ref. 10] in 1964. The questionnaire provides information on such factors as social interaction and effect; speech, motor, and manipulative ability; intelligence and reaction to sensory stimuli; family characteristics; illness development; and physiological data. Questionnaires completed by parents and doctors of mentally disturbed children have been returned to Dr. Rimland as a result of its appearance in the book and from his speaking to clinics and parent groups. At the present time, the completed checklist has been accumulated on approximately 2225 children.

A method of "scoring" the completed questionnaires has been developed [Ref. 15] which gives indications of the degree to which a child exhibits classic early infantile autism characteristics. One autistic behavior point is accrued for each question characteristic of autistic behavior and one non-autistic behavior point is scored for each question answered in the non-autistic direction. Similarly, scores are obtained for autistic speech and non-autistic speech. These scores can be weighted and combined in various ways to arrive at "rational" scores which depict degree of autism. As part of his doctoral dissertation at the University of California, Berkeley, James R. Cameron transformed the 80 items of the Rimland checklist into a 144-item format and obtained ten factor scores.

For approximately 300 of the children, a 24-week study of vitamin effects on behavior was made and an overall vitamin improvement score for each child was obtained. Reference 16 contains the details and results of this study.

It is hoped that through utilization of various clustering techniques homogeneous groups can be obtained, the vitamin effects of different groups evaluated and possibly additional steps toward determining a reliable treatment can be made.

IV. RESULTS OF CLUSTERING TECHNIQUES

A. PREVIOUS STUDIES

There is no "correct" way to cluster data and a variety of methods are available, each requiring a different set of assumptions and utilizing different aspects of the measurements as the basis for discrimination between groups. Rimland [Ref. 16] gives results of several clustering methods previously applied to the questionnaire data and presents clinical findings from the vitamin treatment studies. The remainder of this section summarizes the findings of these studies and reports the results of the application of the STEPWIZE procedure to data described in the preceding section.

NORMIX is a system of cluster analysis developed by John Wolf [Ref. 17]. This procedure was applied using 17 scores derived from Rimland's questionnaires as the basis for the analysis. Ten of the scores were derived by Cameron from a factor analysis of the 144-item format discussed in Section III and seven of the scores were taken from the set of rational scores developed by Rimland. This analysis produced six subgroups of children; and after classification, the mean vitamin improvement score for each group was determined. These means ranged from 46.71 to 69.00 and were shown through analysis of variance to be significantly different at the .02 level. A summary of the results of this scheme is shown in Table II.

Group number	1	2	3	4	5	6
No. in group	19	5	48	35	7	69
Mean vitamin improvement	64.26	69.00	64.46	69.71	46.71	65.01

TABLE II. RESULTS OF NORMIX ANALYSIS

A second computer cluster analysis of the same data was performed by J. R. Cameron of Napa State Hospital. This method (BC TRY) involves different mathematical assumptions and Cameron chose to use only the ten factor scores in his analysis. He produced eight clusters of children, with mean improvement scores ranging from 57.83 to 78.23. Analysis of variance yielded an F ratio of 2.49 which is significant at the .02 level; again indicating a correlation between the symptoms and the degree of improvement which is too great to be explained by chance. Results of this study are summarized in the following table:

Group number	Number in group	Mean vitamin improvement
1	38	65.92
2	13	78.23
3	18	57.88
4	8	61.25
5	25	65.84
6	34	64.40
7	27	65.29
8	13	59.38

TABLE III. RESULTS OF BC TRY ANALYSIS

A third computer cluster analysis was performed by Paul Hoffman at the Oregon Research Institute. Results of this

analysis have not been obtained but, as reported by Rimland, Hoffman used all 17 scores and produced 14 subgroups (the fourteenth consisted of only one child). For the other 13 groups, mean vitamin improvement scores ranged from 58.83 to 71.22, but analysis of variance on improvement scores yielded an F-ratio of 1.32 which is not significant.

B. RESULTS OF STEPWISE ANALYSIS

The clustering technique described in Section II was applied (using the ten factor scores obtained by Cameron) to 225 of the children for which both questionnaire and vitamin improvement scores were available. These scores, depicting early onset, family education, prematurity, rocking as an infant, stiffness, coordination, retardation, resistance to change, social awareness, and destructiveness had been scaled by Cameron so that each had a mean of about 500 and standard deviation of about 70. The particular factor analysis technique used is not known. However, he stated that he eliminated some of the 144 variables because of low frequency of occurrance and others because of high correlation, and that with these ten scores accounted for approximately sixty per cent of the variance of his data set.

Results from the application of the stepwise technique with these discriminators was disappointing in that very early in the scheme, a trend toward establishing one large cluster was apparent. This probably could have been predicted since this technique considered each of the discriminators as

essentially equally weighted. Two children showing quite dissimilar characteristics for some scores were very similar for others, and an overall average point in Euclidean 10 space was reached early in the scheme. It is felt that the assumption of equal weighting of such a variety of characteristics as those depicted by the factors is not valid. It is possible that if the factor scores were scaled so that the variability of each score was directly related to its relative importance in determining autistic characteristics, the stepwise procedure would yield more meaningful results.

A second clustering attempt using the same technique but different discriminators was made. This analysis utilized the four rational scores depicting autistic behavior, non-autistic behavior, autistic speed and non-autistic speech as the measurement space. As discriminators for determining "degree of autism" the behavior scores are considered to be of much more value. Autistic and non-autistic behavior have means of 17.6 and variance of about 47. Autistic speech and non-autistic speech have means of 5.5 and 1.93 with variances of 9.4 and 5.9 respectively. This hierarchical method yields a pyramidal structure of subgroups ranging from 225 subgroups, with each consisting of one individual, to a point where all children are combined into a single cluster. It is possible that some function of the similarities within and between clusters could have been established to aid in choosing which of these configurations gives the "optimal" arrangement. However, the similarity between the two items combined for any

iteration serves as a relative measure of the similarity used for that iteration. The following table lists iteration number and similarity used for iterations 209 through 216.

Iteration number	Iteration similarity measure
209	.008
210	.010
211	.011
212	.009
213	.016
214	.068
215	.017
216	.018

TABLE IV. SIMILARITY USED FOR ITERATIONS

This relatively high value for iteration 214 indicates that two relatively dissimilar items (in this case, two clusters) were combined and that the configuration near iteration 214 should be investigated. At this point in the scheme, all but four of the 225 individuals had been clustered and there existed eight subgroups. Results after step 213 are given in the following table:

Group number	Number in Group	Mean vitamin improvement
1	35	64.43
2	24	73.08
3	74	64.80
4	66	67.17
5	6	72.50
6	6	56.50
7	5	48.80
8	5	60.40

TABLE V. RESULTS OF STEPWISE TECHNIQUE

As in the three previously mentioned studies, analysis of vitamin improvement scores was made through analysis of variance. This resulted in an F-ratio of 2.503 with 7 and 213 degrees of freedom, which is significant at the .025 level. Also of interest in comparison of the clustering configurations with improvement scores is that iteration 214, with its relatively low similarity, combined clusters five and seven (two clusters whose vitamin improvement scores are quite different). To provide information on the measurement characteristics of the different groups, rational score means are provided in the following table:

Cluster number	Number in cluster	Autistic behavior	Non-autistic behavior	Autistic speech	Non-autistic speech
1	35	24.11	11.32	6.25	1.48
2	24	27.79	5.96	8.40	.75
3	74	18.00	14.51	7.01	1.40
4	66	11.83	23.61	5.70	1.95
5	6	5.66	30.01	4.39	2.18
6	6	15.90	17.33	3.00	5.17
7	5	5.21	29.00	1.00	6.88
8	5	9.99	24.21	.80	8.00
Total	221	17.61	17.63	5.50	1.93

TABLE VI. MEASUREMENT MEANS BY CLUSTER

Cluster 2 (high vitamin improvement) showed high autistic behavior, low non-autistic behavior, high autistic speech, and low non-autistic speech. Some contradiction is provided by cluster 5: These children also showed high vitamin improvement, but exhibited essentially opposite behavior scores,

with speech scores near the mean for the entire group. Cluster 7 also adds to the confoundment. With almost the same behavior scores as cluster 5, but with low autistic speech and high non-autistic speech scores, these children scored very low on vitamin improvement. It should be noted that on the next iteration, cluster five and cluster seven were combined, and that cluster two remained intact until the point in the scheme where only three clusters were distinguished. Cluster two is the only group which showed substantially better improvement scores and which remained intact through most of the iterations.

Rimland [Ref. 15] provides support for the contention that overall autistic scores derived from the questionnaire can differentiate early infantile autism. The overall score is obtained simply by summing the two autistic scores and subtracting the non-autistic scores. An overall score of +20 or higher is regarded as highly indicative of early infantile autism and only about 9.7 per cent of the entire sample reach this score. Using the mean rational scores for the groups resulting from the STEPWISE cluster analysis, only group two displays characteristics highly indicative of early infantile autism.

Due to the small number of individuals in some groups and the near mean vitamin improvement for others, little additional information as to the "type" of individuals significantly helped by the vitamin dosages can be obtained. Tests for significantly different vitamin improvement means

between groups one and four, for example, yielded an F-ratio of only .889 which is not significant.

Major William Knauer, USA, conducted a second factor analysis of the 144-variable data set. He eliminated 65 variables because of high correlation or low frequency of occurrance and obtained ten factor scores. These ten factors were not scaled when used as the measurement space for the STEPWISE clustering technique so the "natural" variance associated with the factors dictated their relative importance in computing Euclidean distances. The results of this attempt were similar to the attempt using Cameron's factors in that at points in the scheme of greatest interest (when most of the individuals had been clustered), the individuals tended to accumulate into one large cluster. After iteration number 196, there existed eight clusters and 203 of the individuals had been clustered. One cluster contained 179, one contained 10, two contained three, and four clusters had only two individuals each.

The STEPWISE procedure was applied to three different sets of measurements all taken from the questionnaire data, and only Rimland's rational scores provided groups which can be investigated as to their homogeneity and the relative effects of vitamin dosages.

V. CONCLUSIONS

The application of a particular clustering scheme to a particular set of data involves assumptions about the appropriateness of the statistical and mathematical techniques employed in the scheme. These assumptions are often difficult to justify and the researcher must rely to some extent on intuition and experience with the characteristics of the objects under consideration; it would certainly be unwise to accept the results of the first scheme applied. The results discussed in this paper were all obtained from clustering based on factor scores and/or rational scores taken from the questionnaire data. It is possible that other symptoms or additional measurements would be appropriate as discriminators; and results of various clustering schemes, with their different assumptions, should be compared in order to obtain more credible information.

Although three different clustering techniques have produced groups whose mean vitamin improvement scores were found to be significantly different, the difference can be attributed to a relatively small number of individuals. Different discriminators, and in the case of STEPWISE procedure, a slightly different group of children, were used for the clustering. However, in each case there seems to be one group which displays considerably better improvement scores. The NORMIX analysis had one group of 35 children

whose mean improvement score was 69.7. The other five groups either had very few individuals or average improvement scores near the overall mean. The BC TRY method resulted in a group of 13 children with a mean score of 78.23 and the STEPWISE procedure had 24 in its "high group" (four of which were not used with the other two methods). The "high improvement groups" resulting from these three procedures contained 35, 13, and 20 individuals, with only three children common to all three groups. An additional six were common to the NORMIX and BC TRY groups, four to NORMIX and STEPWISE, and one common to BC TRY and STEPWISE. The BC TRY "low improvement" group contained 18 individuals, NORMIX seven, and the STEPWISE had five (one of which was not used with the other methods). None of the 18 individuals in the BC TRY low group were in either of the other two low groups, and two children were common to both NORMIX and STEPWISE.

As the above discussion points out, the three methods produced groups whose improvement scores were different, but they each produced somewhat different sets of "homogeneous" groupings. Little information about the types of children who are likely to be helped can be extracted. However, there seems to be some credibility added to the contention of behavioral improvements as a result of vitamin dosages.

As discussed in Section IV, the STEPWISE technique produced one group of 24 children whose mean improvement score was substantially higher than the others and this group remained intact until only two clusters were distinguished.

It is of some possible significance that with Rimland's method of determining degree of autism, this is the only group which would be considered to be highly indicative of early infantile autism.

The disappointing results from the application of the STEPWISE technique to factor scores certainly does not mean that factor analysis is the wrong approach; however, some general questions are raised: The responses from the 80-item questionnaire were transformed into a 144-item format which must have induced some added correlation. Some of the 144 variables were then eliminated because of low frequency or high correlation, and then factor analysis, which considers the dependence structure, was applied. Intuitively it seems that some information originally contained in the 80 items could have been lost or altered through this procedure. A factor analysis of the original 80 items or a statistically elegant clustering scheme using the 80 scores as measurements are suggested as possible topics for future statistical studies of the current data.

As previously mentioned, the qualitative nature of this data would complicate the clustering problem. A series of qualitative measurements, e.g., one with two states, one with three states and one with four states, defines 24 multivariate states and the responses an individual makes to the measurements defines which of the 24 states the individual occupies. Determining interstate similarities could be accomplished only by a thorough study of the intent of the

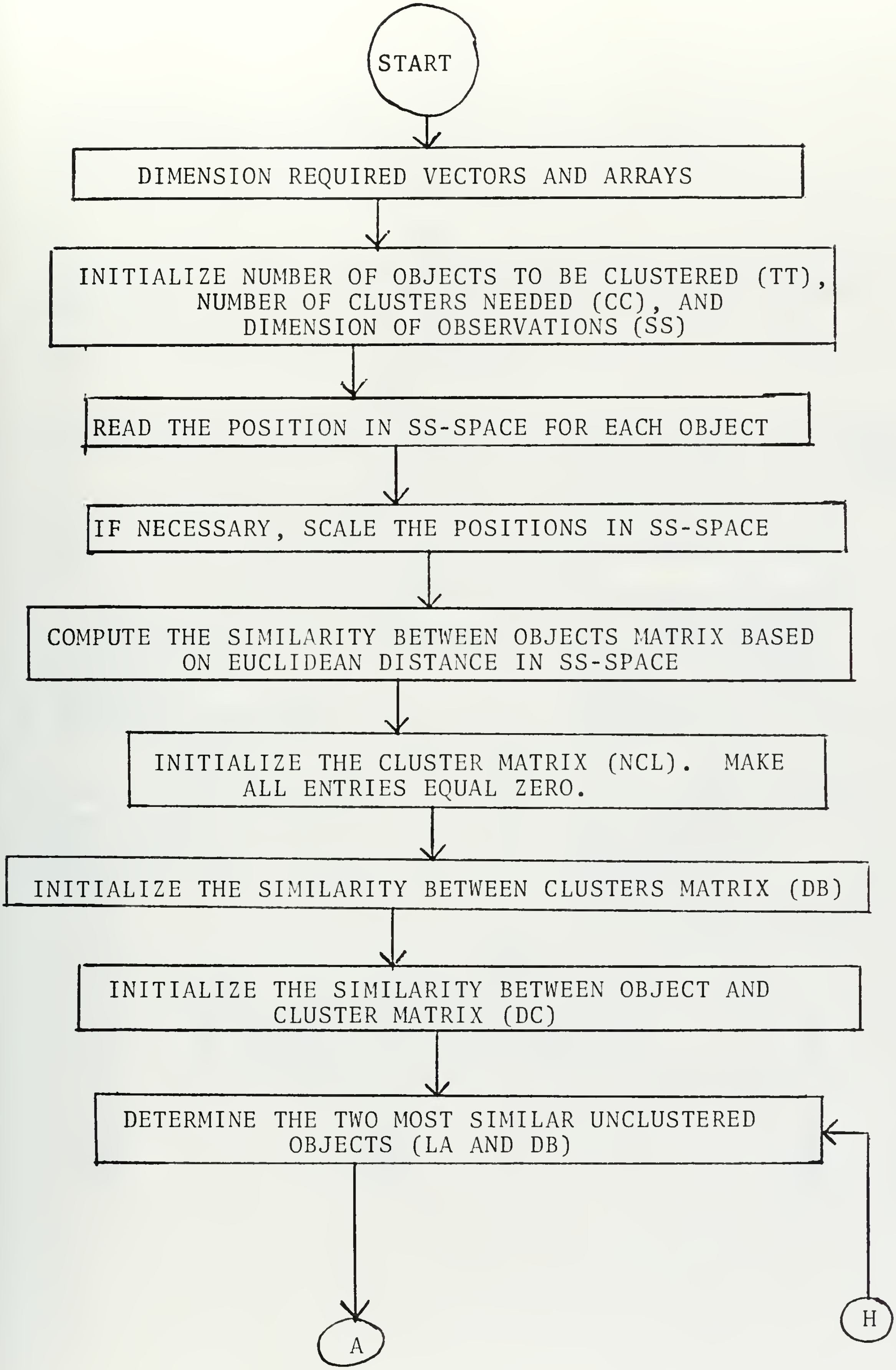
measurements. However, if some importance ranking of the measurements and their response states can be obtained, a clustering technique based on occupancy of the qualitative multivariate states can be developed.

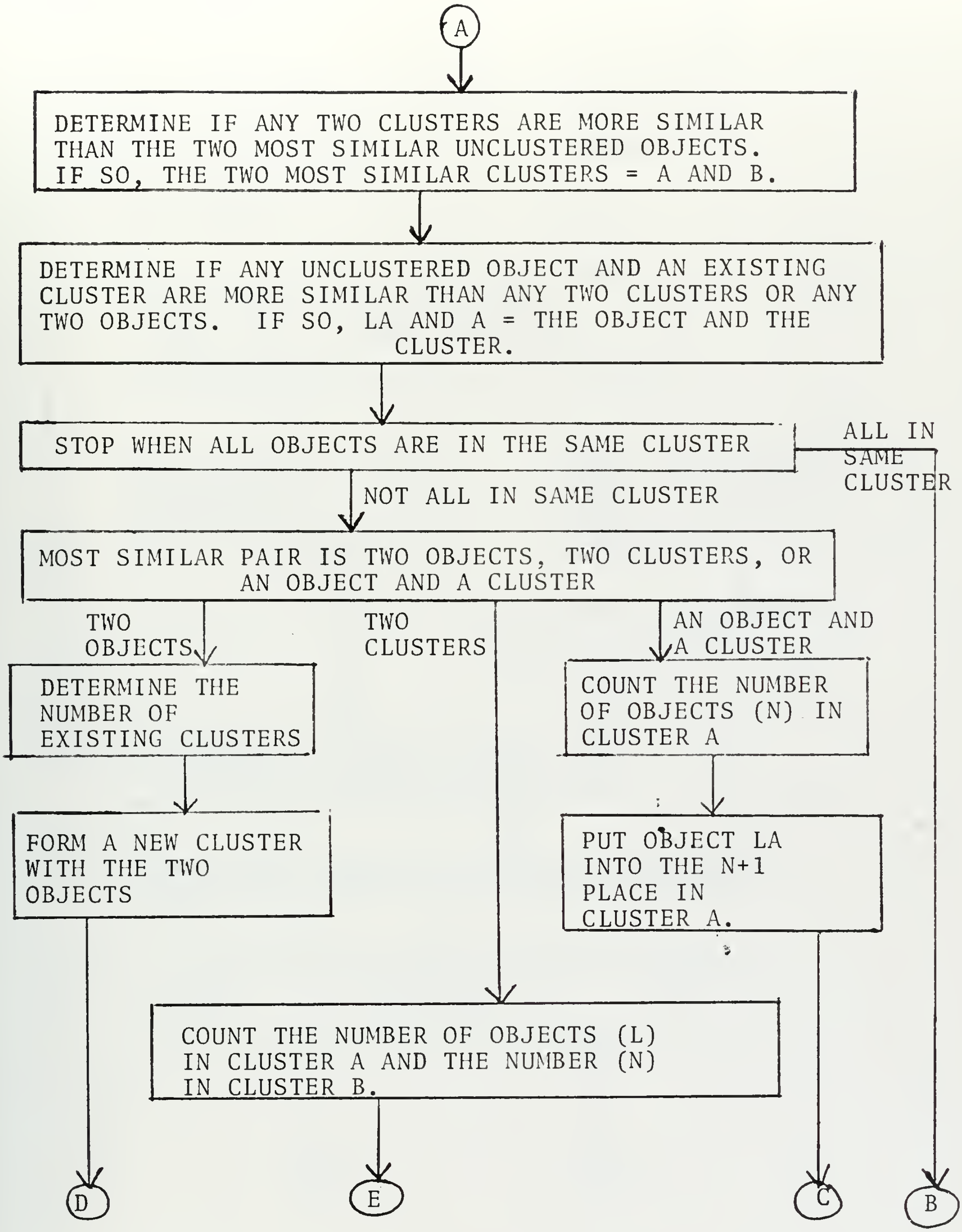
No conclusive comments about the comparative performance of various clustering techniques can be made on the basis of performance with one data set. Each technique involves different assumptions about the appropriateness of the data, the relative importance given to the different measurements, and the means of obtaining the discriminating measurement variables. Comparisons are also complicated by the fact that the "correct" solution to the problem is not known. The STEPWISE procedure, when applied to the rational scores, produced results which were "similar" to those produced by other methods, in that eight separate groups were distinguished and statistical tests on vitamin improvements produced similar results. It is felt that these results provide some justification for the straightforward approach offered by hierarchical procedures.

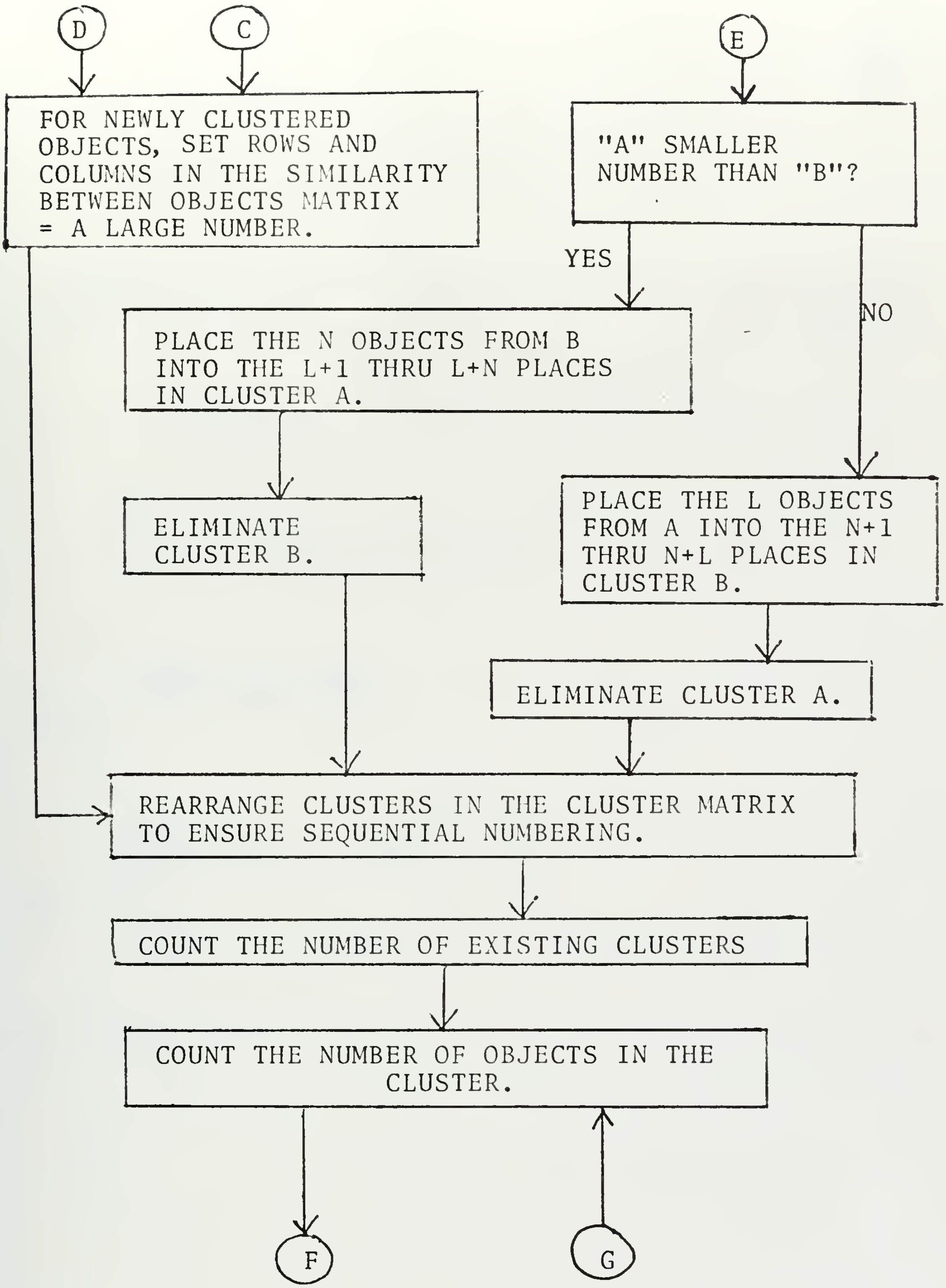
Due to the uncertainty regarding the adequacy of the selection of the measurements depicting symptoms which distinguish various forms of child psychosis, and the inherent difficulties associated with statistical clustering, no strong claim can be made to the methods' reliabilities. It is believed that the results obtained by application of the STEPWISE procedure and by the previously conducted studies offered some credibility that vitamin treatments can result

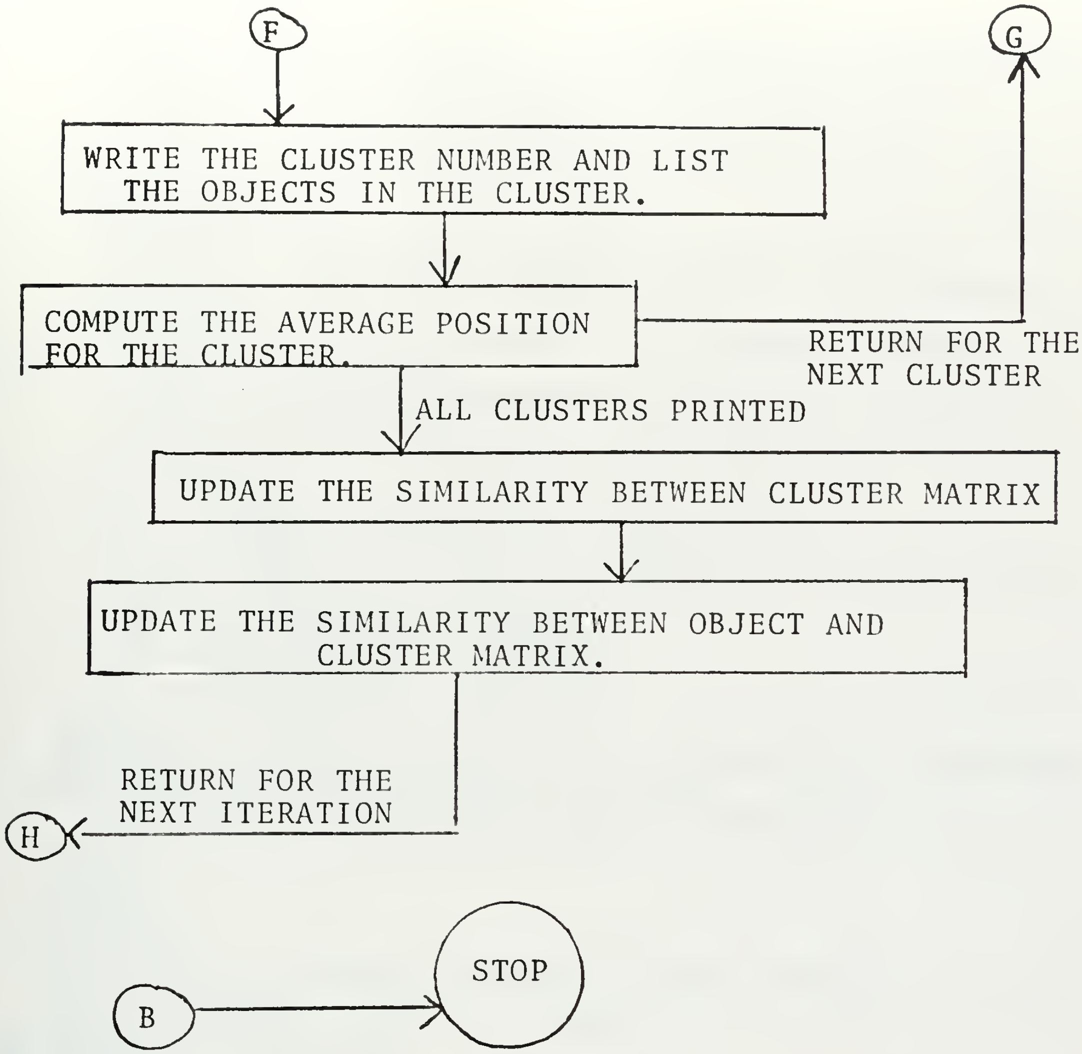
in behavioral improvements for certain types of psychotic children. It is hoped that through the efforts of Dr. Rimland, and others in the field, significant breakthroughs in the etiology and treatment of child psychosis can be accomplished.

APPENDIX A


START
DIMENSION REQUIRED VECTORS AND ARRAYS
INITIALIZE NUMBER OF OBJECTS TO BE CLUSTERED (TT), NUMBER OF CLUSTERS NEEDED (CC), AND DIMENSION OF OBSERVATIONS (SS)
READ THE POSITION IN SS-SPACE FOR EACH OBJECT
IF NECESSARY, SCALE THE POSITIONS IN SS-SPACE
COMPUTE THE SIMILARITY BETWEEN OBJECTS MATRIX BASED ON EUCLIDEAN DISTANCE IN SS-SPACE
INITIALIZE THE CLUSTER MATRIX (NCL). MAKE ALL ENTRIES EQUAL ZERO.
INITIALIZE THE SIMILARITY BETWEEN CLUSTERS MATRIX (DB)
INITIALIZE THE SIMILARITY BETWEEN OBJECT AND CLUSTER MATRIX (DC)
DETERMINE THE TWO MOST SIMILAR UNCLUSTERED OBJECTS (LA AND DB)
A
H


A
DETERMINE IF ANY TWO CLUSTERS ARE MORE SIMILAR THAN THE TWO MOST SIMILAR UNCLUSTERED OBJECTS. IF SO, THE TWO MOST SIMILAR CLUSTERS = A AND B.
DETERMINE IF ANY UNCLUSTERED OBJECT AND AN EXISTING CLUSTER ARE MORE SIMILAR THAN ANY TWO CLUSTERS OR ANY TWO OBJECTS. IF SO, LA AND A = THE OBJECT AND THE CLUSTER.
STOP WHEN ALL OBJECTS ARE IN THE SAME CLUSTER
ALL IN SAME CLUSTER
NOT ALL IN SAME CLUSTER
MOST SIMILAR PAIR IS TWO OBJECTS, TWO CLUSTERS, OR AN OBJECT AND A CLUSTER
TWO OBJECTS
TWO CLUSTERS
AN OBJECT AND A CLUSTER
DETERMINE THE NUMBER OF EXISTING CLUSTERS
COUNT THE NUMBER OF OBJECTS (N) IN CLUSTER A
FORM A NEW CLUSTER WITH THE TWO OBJECTS
PUT OBJECT LA INTO THE N+1 PLACE IN CLUSTER A.
COUNT THE NUMBER OF OBJECTS (L) IN CLUSTER A AND THE NUMBER (N) IN CLUSTER B.
D
E
C
B


D
C
E
FOR NEWLY CLUSTERED OBJECTS, SET ROWS AND COLUMNS IN THE SIMILARITY BETWEEN OBJECTS MATRIX = A LARGE NUMBER.
"A" SMALLER NUMBER THAN "B"?
YES
NO
PLACE THE N OBJECTS FROM B INTO THE L+1 THRU L+N PLACES IN CLUSTER A.
ELIMINATE CLUSTER B.
PLACE THE L OBJECTS FROM A INTO THE N+1 THRU N+L PLACES IN CLUSTER B.
ELIMINATE CLUSTER A.
REARRANGE CLUSTERS IN THE CLUSTER MATRIX TO ENSURE SEQUENTIAL NUMBERING.
COUNT THE NUMBER OF EXISTING CLUSTERS
COUNT THE NUMBER OF OBJECTS IN THE CLUSTER.
F
G


F
G
WRITE THE CLUSTER NUMBER AND LIST THE OBJECTS IN THE CLUSTER.
COMPUTE THE AVERAGE POSITION FOR THE CLUSTER.
RETURN FOR THE NEXT CLUSTER
ALL CLUSTERS PRINTED
UPDATE THE SIMILARITY BETWEEN CLUSTER MATRIX
UPDATE THE SIMILARITY BETWEEN OBJECT AND CLUSTER MATRIX.
RETURN FOR THE NEXT ITERATION
H
B
STOP

```
C
C
C
C        THIS PROGRAM PROVIDES A CLUSTERING SCHEME BASED ON
C        SIMILARITY MATRICES BETWEEN OBJECTS, BETWEEN
C        CLUSTERS AND BETWEEN OBJECTS AND CLUSTERS.
C        * * * * * * * * * LEGEND * * * * * * * * * * *
C        D(I,J)---SIMILARITY BETWEEN OBJECTS MATRIX
C        XND(I,J)-MATRIX GIVING POSITION IN SS-SPACE FOR
C        EACE OBJECT.
C        NCL(I,J)-MATRIX WHICH CONTAINS THE OBJECTS IN EACH
C        CLUSTER
C        DB(I,J)--SIMILARITY BETWEEN CLUSTERS MARTIX.
C        DC(I,J)--SIMILARITY BETWEEN OBJECTS AND CLUSTERS
C        MATRIX.
C        XSM(I)---WORKING VECTOR FOR COMPUTING CLUSTER
C        AVERAGE POSITIONS.
C        XSU(I)---WORKING VECTOR FOR COMPUTING CLUSTER
C        AVERAGE POSITIONS.
C        XST(I,J)-MATRIX GIVING AVERAGE POSITION FOR EACH
C        CLUSTER.
C        TT---NUMBER OF OBJECTS TO BE CLUSTERED.
C        CC---NUMBER OF CLUSTERS NEEDED.
C        SS---DEMENSIONALITY OF OBSERVATIONS.
C        DIST-SIMILARITY BETWEEN OBJECTS OR CLUSTERS WHICH
C        WERE COMBINED ON THE ITTERATION.
C        LA---DESIGNATES AN OBJECT.
C        LB---DESIGNATES AN OBJECT.
C        A----DESIGNATES A CLUSTER.
C        B----DESIGNATES A CLUSTER.
C     * * * * * * * * * * * * * * * * * * * * * * * * * * *
      IMPLICIT INTEGER (A,B,C,T,S)
      DIMENSION D(225,225),XND(225,4),NCL(60,225),DB(60,60),
     1DC(225,60),XSM(4),XSU(4),XST(60,4)
C        SET THE NUMBER OF OBJECTS (TT),CLUSTERS (CC), AND
C        THE MEASUREMENT SPACE (SS).
      TT=225
      CC=60
      SS=4
      TC=TT-1
      CT=CC-1
C        READ THE SS-SPACE POINT FOR EACH OBJECT.
      DO 801 I=1,TT
      READ (5,800) (XND(I,J),J=1,SS)
  800 FORMAT (7X,10F7.3)
  801 CONTINUE
C        SCALE ALL SCORES SO THAT SIMILARITY BETWEEN OBJECTS
C        DOES NOT EXCEED 88880.
      DO 5320 I=1,TT
      DO 5321 J=1,SS
      XND(I,J)=XND(I,J)/10.
 5321 CONTINUE
 5320 CONTINUE
C        COMPUTE THE SIMILARITY MATRIX (D) BASED ON WEIGHTED
C        EUCLIDEAN SS-SPACE.
      DO 803 I=1,TT
      DO 804 J=1,TT
      YA=0.
      DO 802 K=1,SS
      X=XND(I,K)
      Y=XND(J,K)
      XA=(X-Y)**2
      YA=YA+XA
  802 CONTINUE
      D(I,J)=YA
      D(J,I)=YA
  804 CONTINUE
  803 CONTINUE
      WRITE (6,3300)
 3300 FORMAT ('1',15X,'OUTPUT')
C        MAKE THE CLUSTER MATRIX EQUAL ZERO.
      DO 2600 I=1,CC
```

```
      DO 2700 J=1,TT
      NCL(I,J)=0
 2700 CONTINUE
 2600 CONTINUE
C        MAKE THE SIMILARITY BETWEEN CLUSTER MATRIX LARGE.
      DO 5020 I=1,CC
      DO 5021 J=1,CC
      DB(I,J)=88888.
      DB(J,I)=88888.
 5021 CONTINUE
 5020 CONTINUE
C        MAKE THE SIMILARITY BETWEEN OBJECT AND CLUSTER
C        MATRIX LARGE.
      DO 5022 I=1,TT
      DO 5023 J=1,CC
      DC(I,J)=88888.
 5023 CONTINUE
 5022 CONTINUE
C        FIND THE TWO MOST SIMILAR UNCLUSTERED OBJECTS
C        (LA,LB).
      NP=1
 1111 DIST=88887.
      LAB=1
      DO 100 I=1,TC
      K=I+1
      IF (D(I,1).GE.88885.) GO TO 100
      DO 200 J=K,TT
      IF (D(J,1).GE.88885.) GO TO 200
      IF (D(I,J).GE.DIST) GO TO 200
      LA=I
      LB=J
      DIST=D(I,J)
      LAB=0
  200 CONTINUE
  100 CONTINUE
C        IF ANY TWO OF THE CLUSTERS ARE MORE SIMILAR THAN
C        THE TWO MOST SIMILAR OBJECTS FIND THESE TWO CLUSTER
C        (A,B).
      DO 5000 I=1,CT
      K=I+1
      DO 5001 J=K,CC
      IF (DB(I,J).GE.DIST) GO TO 5001
      A=I
      B=J
      DIST=DB(I,J)
      LAB=2
 5001 CONTINUE
 5000 CONTINUE
C        IF THE SIMILARITY BETWEEN AN OBJECT AND A CLUSTER
C        IS MORE SIMILAR THAN THE TWC CLUSTERS OR THE TWO
C        OBJECTS FIND THE OBJECT AND CLUSTER (LB,A).
      DO 5010 I=1,TT
      DO 5011 J=1,CC
      IF (D(I,1).GE.88885.) GO TO 5010
      IF (DC(I,J).GE.DIST) GO TO 5011
      LB=I
      LA=I
      LAA=1
      A=J
      LAB=1
      DIST=DC(I,J)
 5011 CONTINUE
 5010 CONTINUE
C        IF EVERY OBJECT IS IN THE SAME CLUSTER--STOP
      IF (DIST.GE.88880.) GO TO 2223
      NP=NP+1
C        IF FORMING A NEW CLUSTER GO TO 460.  IF COMBINING
C        TWO CLUSTERS GO TO 470.
      IF (LAB.EQ.0) GO TO 460
      IF (LAB.EQ.2) GO TO 470
C        COUNT THE NUMBER (N) OF OBJECTS IN CLUSTER A AND
C        PUT OBJECT LB INTO THE N+1 PLACE.
```

```
      N=0
      DO 430 I=1,NP
      IF (NCL(A,I).NE.0) N=N+1
      IF(N.LT.I) GO TO 435
  430 CONTINUE
  435 N=N+1
      NCL(A,N)=LB
      GO TO 9999
C        FORM A NEW CLUSTER.  PUT OBJECTS LA AND LB INTO THE
C        FIRST TWO PLACES.
  460 DO 500 I=1,NP
      IF (NCL(I,1).NE.0) GO TO 500
      NCL(I,1)=LA
      NCL(I,2)=LB
      GO TO 9999
  500 CONTINUE
C        JOIN THE TWO CLUSTERS A AND B. COUNT THE NUMBER (L)
C        OF OBJECTS IN CLUSTER A AND THE NUMBER (N) IN
C        CLUSTER B.
  470 L=0
      DO 600 J=1,NP
      IF (NCL(A,J).NE.0) L=L+1
      IF (L.LT.J) GO TO 620
  600 CONTINUE
  620 N=0
      DO 650 J=1,NP
      IF (NCL(B,J).NE.0) N=N+1
      IF (N.LT.J) GO TO 630
  650 CONTINUE
C        IF A LESS THAN B PUT THE N OBJECTS FROM B INTO THE
C        L+1 THRU L+N PLACES IN CLUSTER A.  ELIMINATE B.
  630 M=L+1
      K=N+1
      J=L+N
      IF (B.LT.A) GO TO 700
      C=1
      DO 680 I=M,J
      NCL(A,I)=NCL(B,C)
      NCL(B,C)=0
      C=C+1
  680 CONTINUE
      GO TO 9998
C        IF B LESS THAN A PUT THE L OBJECTS FROM A INTO THE
C        N+1 THRU N+L PLACES IN CLUSTER B. ELIMINATE A.
  700 C=1
      DO 710 I=K,J
      NCL(B,I)=NCL(A,C)
      NCL(A,C)=0
      C=C+1
  710 CONTINUE
      GO TO 9998
C        MAKE THE ROWS AND COLUMNS (FOR NEWLY CLUSTERED
C        OBJECTS) IN MATRIX D EQUAL A LARGE NUMBER.
 9999 DO 5030 I=1,TT
      J=LB
      K=LA
      D(K,I)=88888.
      D(J,I)=88888.
      IF ((J.EQ.1).AND.(K.EQ.1)) GO TO 5030
      IF (J.EQ.1) GO TO 4998
      IF (K.EQ.1) GO TO 4999
      D(I,K)=88888.
      D(I,J)=88888.
      GO TO 5030
 4998 D(I,K)=88888.
      GO TO 5030
 4999 D(I,J)=88888.
 5030 CONTINUE
C        REARRANGE CLUSTERS TO HAVE SEQUENTIAL NUMBERING.
 9998 DO 3021 I=1,CT
      K=I+1
      IF (NCL(I,1).NE.0) GO TO 3021
```

```
      DO 3022 J=1,TT
      NCL(I,J)=NCL(K,J)
      NCL(K,J)=0
 3022 CONTINUE
 3021 CONTINUE
C        COUNT THE NUMBER OF EXISTING CLUSTERS (K).
      K=0
      DO 3023 I=1,CC
      IF (NCL(I,1).NE.0) K=K+1
      IF (K.LT.I)  GO TO 3222
 3023 CONTINUE
 3222 IF (K.EQ.CC) GO TO 2224
      NPAS=NP-1
      WRITE (6,3301) NPAS
 3301 FORMAT (' ',2X,'CLUSTERS AFTER PASS NUMBER',5X,I5)
C        COUNT THE NUMBER OF OBJECTS IN THE CLUSTER UNDER
C        CONSIDERATION.
      DO 2111 I=1,K
      L=0
      DO 3024 M=1,TT
      IF (NCL(I,M).NE.0) L=L+1
      IF (L.LT.M) GO TO 2199
 3024 CONTINUE
 2199 WRITE (6,2200) I,(NCL(I,J),J=1,L)
 2200 FORMAT (' ',3X,I4,4X,12(20I5,/,' ',11X))
C        COMPUTE THE AVERAGE POSITION FOR EACH CLUSTER.
      DO 4003 N=1,SS
      XSM(N)=0.
 4003 CONTINUE
      DO 4001 B=1,L
      A=NCL(I,B)
      DO 4000 N=1,SS
      XSU(N)=XND(A,N)
      XSM(N)=XSM(N)+XSU(N)
 4000 CONTINUE
 4001 CONTINUE
      DO 4005 N=1,SS
      XL=L
      XST(I,N)=XSM(N)/XL
 4005 CONTINUE
 2111 CONTINUE
      WRITE (6,5328) DIST
 5328 FORMAT (' ','DIST=',2X,F10.3)
C        RECOMPUTE THE SIMILARITY BETWEEN CLUSTER MATRIX.
      DO 4010 B=1,K
      DO 4011 C=1,K
      XA=0.
      DO 4012 T=1,SS
      Y=XST(B,T)
      Z=XST(C,T)
      YA=(Y-Z)**2
      XA=XA+YA
 4012 CONTINUE
      DB(B,C)=XA
      DB(C,B)=XA
 4011 CONTINUE
 4010 CONTINUE
C        MAKE THE ROWS AND COLUMNS (FOR ELIMINATED
C        CLUSTERS) IN THE DB MATRIX LARGE.
      NK=K+1
      DO 4013 B=1,CC
      DO 4014 C=NK,CC
      DB(C,B)=88888.
      DB(B,C)=88888.
 4014 CONTINUE
 4013 CONTINUE
C        UPDATE THE DC MATRIX.
      DO 4015 B=1,TT
      IF (D(B,1).GE.88885.) GO TO 4015
      DO 4016 C=1,K
      XA=0.
      DO 4017 T=1,SS
```

```
      Y=XND(B,T)
      Z=XST (C,T)
      YA=(Y-Z)**2
      XA=XA+YA
 4017 CONTINUE
      DC(B,C)=XA
 4016 CONTINUE
 4015 CONTINUE
C        GO BACK FOR ANOTHER ITTERATION.
      WRITE (6,5329)
 5329 FORMAT (' ','COMPLETE PASS')
      GO TO 1111
 2224 WRITE (6,2225)
 2225 FORMAT (' ','TOO MANY CLUSTERS')
 2223 FIN=3.0
      WRITE (6,5555)
 5555 FORMAT (' ','HAD A NORMAL ENDING')
      STOP
      END
```

APPENDIX C

1. Present Age of Child:

    1. Under 3 years old
    2. Between 3 and 4 years old
    3. Between 4 and 5 years old
    4. Between 5 and 6 years old
    5 Over 6 years old

2. Child's Sex:

    1. Boy
    2. Girl

3. Child's Birth Order and Number of Mother's Other Children:

    1. Only child
    2. First born
    3. Last born
    4. Middle born
    5. Foster children

4. Were Pregnancy and Delivery Normal?

    1. Pregnancy and delivery both normal
    2. Problems during both pregnancy and delivery
    3. Pregnancy troubled, routine delivery
    4. Pregnancy untroubled, problems during delivery
    5. Don't know

5. Was the Birth Premature (Birth Weight under 5 lbs.)?

    1. Yes
    2. No
    3. Don't know

6. Was the Child Given Oxygen in the First Week?

    1. Yes
    2. No
    3. Don't know

7. Appearance of Child during First Few Weeks after Birth:

    1. Pale, delicate looking
    2. Unusually healthy looking
    3. Average, don't know, or other

8. Unusual Conditions of Birth and Infancy:

   1. Unusual conditions, including blindness, birth ingury
   2. Twin birth
   3. Both one and two
   4. Normal, or don't know

9. Baby's Health in First Three Months:

   1. Excellent health, no problems
   2. Respiration problems
   3. Skin problems
   4. Feeding problems
   5. Elimination problems
   6. Several of the above

10. Has the Child Been Given an EEG?

   1. Yes, normal
   2. Yes, borderline
   3. Yes, abnormal
   4. No, don't know or don't know results

11. Reactions to Bright Lights and Colors, Unusual Sounds, etc., during First year:

   1. Unusual strong reaction
   2. Unusually unresponsive
   3. Average or don't know

12. Did the Child Behave Normally for a Time before His Abnormal Behavior Began?

   1. Never was a period of normal behavior
   2. Normal during first six months
   3. Normal during first year
   4. Normal during first 1-1/2 years
   5. Normal during first 2 years
   6. Normal during first 3 years
   7. Normal during first 4-5 years

13. (Age 4-8 Mo.) Did the Child Reach Out or Prepare Himself to be Picked Up when Mother Approached Him?

   1. Yes, or I believe so
   2. No, I don't think he did
   3. No, definitely not
   4. Don't know

14. Did the Child Rock in his Crib as a Baby?

   1. Yes, quite a lot
   2. Yes, sometimes
   3. No, or very little
   4. Don't know

15. At What Age did the Children Learn to Walk Alone?

   1. 8-12 mo.
   2. 13-15 mo.
   3. 16-18 mo.
   4. 19-24 mo.
   5. 25-30 mo.
   6. 37 mo. or later, or does not walk alone

16. Which Describes the Change from Crawling to Walking?

   1. Normal change from crawling to walking
   2. Little or no crawling, gradual start of walking
   3. Little or no crawling, sudden start of walking
   4. Prolonged crawling, sudden start of walking
   5. Prolonged crawling, gradual start of walking
   6. Other, or don't know

17. During the Child's First Year, Did he Seem to be Unusually Intelligent?

   1. Suspected high intelligence
   2. Suspected average intelligence
   3. Child looked somewhat dull

18. During the Child's First Two Years, Did He Like to be Held?

   1. Liked being picked up; enjoyed being held
   2. Limp and passive on being held
   3. You could pick up and hold child only as he wished
   4. Notably stiff and awkward to hold
   5. Don't know

19. Before Age 3, Did the Child Ever Imitate Another Person?

   1. Yes, waved bye-bye
   2. Yes, played pat-a-cake
   3. Yes, other
   4. Two or more of the above
   5. No, or not sure

20. Before Age 3, Did the Child Have an Unusually Good Memory?

   1. Remarkable memory for songs, rhymes, T.V. commercials
   2. Remarkable memory for songs, music (humming only)
   3. Remarkable memory for names, places, routes, etc.
   4. No evidence for remarkable memory
   5. Apparently rather poor memory
   6. Both 1 and 3
   7. Both 2 and 3

21. Did You Ever Suspect the Child Was Very Nearly Deaf?

1. Yes
2. No

22. (Age 2-4) Is Child "Deaf" to Some Sounds but Hears Others?

1. Yes, can be "deaf" to loud sounds, but hears low ones
2. No, this is not true of him.

23. (Age 2-4) Does the Child Hold His Hands in Strange Postures?

1. Yes, sometimes or often
2. No

24. (Age 2-4) Does Child Engage in Rhythmic or Rocking Activity for Very Long Periods of Time (Like on Rocking-Horse or Chair)?

1. Yes, this is typical
2. Seldom does this
3. Not true of him

25. (Ages 2-4) Does the Child Ever "Look Through" or "Walk Through" People?

1. Yes, often
2. Yes, I think so
3. No, doesn't do this

26. (Ages 2-5) Does the Child Have any Unusual Cravings for Things to Eat or Chew on?

1. Yes, salt or salty food
2. Yes, often chews metal objects
3. Yes, other
4. Yes, more than two above
5. No, or not sure

27. (Ages 2-4) Does the Child Have Certain Eating Oddities?

1. Yes, definitely
2. No, or not to any marked degree
3. Don't know

28. Could Child around Ages 3 or 4 be Described as Being "In a Shell" or so Distant and "Lost in Thought" that You Couldn't Reach Him?

1. Yes, this is a very accurate description
2. Once in awhile he might possibly be that way
3. Not an accurate description

29. (Ages 2-5) Is He Cuddly?

    1. Definitely, likes to cling to adults
    2. Above average (likes to be held)
    3. No, rather stiff and awkward to hold
    4. Don't know

30. (Ages 3-5) Does the Child Deliberately Hit His Own Head?

    1. Never, or rarely
    2. Yes, usually by slapping it with his hand
    3. Yes, usually by banging it against another's legs or head
    4. Yes, usually by hitting walls, floor furniture
    5. Several of above

31. (Ages 3-5) How Well Physically Coordinated is the Child?

    1. Unusually graceful
    2. About average
    3. Somewhat below average, or poor

32. (Ages 3-5) Does the Child Sometimes Whirl Himself Like a Top?

    1. Yes, does this often
    2. Yes, sometimes
    3. Yes, if you start him out
    4. No, he shows no tendency to whril

33. (Ages 3-5) How Skillful is the Child in Doing Fine Work with His Fingers or Playing with Small Objects?

    1. Exceptionally skillful
    2. Average for age
    3. A little awkward, or very awkward
    4. Don't know

34. (Ages 3-5) Does the Child Like to Spin Things like Jar Lids, Coins, etc.?

    1. Yes, often, and for rather long periods
    2. Very seldom, or never

35. (Ages 3-5) Does the Child Show an Unusual Degree of Skill at:

    1. Assembling jig-saw or similar puzzles
    2. Arithmetic computations
    3. Can tell day of week a certain date will fall on
    4. Perfect musical pitch

5. Throwing and/or catching ball
6. Other
7. More than one of above
8. No unusual skill, or not sure

36. (Ages 3-5) Does the Child Sometimes Jump Up and Down Gleefully when Pleased?

    1. Yes, this is typical
    2. No, or rarely

37. (Ages 3-5) Does the Child Sometimes Line Things Up in Precise Evenly-Spaced Rows and Insist They Not Be Disturbed?

    1. No
    2. Yes
    3. Not sure

38. (Ages 3-5) Does the Child Refuse to Use His Hands for an Extended Period of Time?

    1. Yes
    2. No

39. Was There a Time before Age Five when the Child Strongly Insisted on Listening to Music on Records?

    1. Yes, insisted on only certain records
    2. Yes, but almost any record would do
    3. Liked to listen, but didn't demand to
    4. No special interest in records

40. (Ages 3-5) How Interested is the Child in Mechanical Objects , the Stove or Vacuum Cleaner?

    1. Little or no interest
    2. Average interest
    3. Fascinated by certain mechanical things

41. (Ages 3-5) How Does the Child Usually React to Being Interrupted at What He is Doing?

    1. Rarely or never gets upset
    2. Sometimes gets mildly upset; rarely very upset
    3. Typically gets very upset

42. (Ages 3-5) Will the Child Readily Accept New Articles of Clothing (Shoes, etc.)?

    1. Usually resists new clothes
    2. Doesn't seem to mind, or enjoys them

43. (Ages 3-5) Is the Child Upset by Certain Things that are Not Right?

1. Not especially
2. Yes, such things often upset him greatly
3. Not sure

44. (Ages 3-5) Does the Child Adopt Complicated "Rituals" which Make Him Very Upset if not Followed?

1. Yes, definitely
2. Not sure
3. No

45. (Ages 3-5) Does Child Get Very Upset if Certain Things He is Used to Are Changed?

1. No
2. Yes, definitely
3. Slightly true

46. (Ages 3-5) Is the Child Destructive?

1. Yes, this is definitely a problem
2. Not deliberately or severely destructive
3. Not especially destructive

47. (Age 3-5) Is the Child Unusually Physically Pliable?

1. Yes
2. Seems normal in this way
3. Definitely not pliable

48. (Age 3-5) Which Single Description or Combination of Descriptions Best Characterizes the Child?

1. Hyperactive, constantly moving
2. Watches television quietly for long periods
3. Sits for long periods, stares into space, or play respectively
4. Combination of 1 and 2
5. Combination of 2 and 3
6. Combination of 1 and 3

49. (Age 3-5) Does the Child Seem to Want to be Liked?

1. Yes, unusually so
2. Just normally so
3. Indifferent to being liked; happiest when left alone

50. (Ages 3-5) Is the Child Sensitive and/or Affectionate?

1. Is sensitive to criticism and affectionate
2. Is sensitive to criticism, not affectionate

3. Not sensitive to criticism, is affectionate
4. Not sensitive to criticism, nor affectionate

51. (Age 3-5) Is it Possible to Direct Child's Attention to an Object Some Distance Away or Out a Window?

1. Yes, no special problem
2. He rarely sees things very far out of reach
3. He examines things with fingers and mouth only

52. (Age 3-5) Do People Consider the Child Especially Attractive?

1. Yes, very good-looking child
2. No, just average
3. Faulty in physical appearance

53. (Age 3-5) Does the Child Look Up at People when They are Talking to Him?

1. Never, or rarely
2. Only with parents
3. Usually does

54. (Age 3-5) Does the Child Take an Adult by the Wrist to Use Adults' Hands?

1. Yes, this is typical
2. Perhaps, or rarely
3. No.

55. (Age 3-5) Which Set of Terms Best Describes the Child?

1. Confused, self-concerned, perplexed, dependent worried
2. Aloof, indifferent, self-contented, remote

56. (Age 3-5) Is the Child Extremely Fearful?

1. Yes, of strangers and certain people
2. Yes, of certain animals, noises or objects
3. Yes, of oneand the above
4. Only normal fearfulness
5. Seems unusually bold and free of fear
6. Child ignores or is unaware of fearsome objects

57. (Age 3-5) Does he Fall or Get Hurt in Running or Climbing?

1. Tends toward falling or injury
2. Average in this way
3. Never, or almost never, exposes self to falling
4. Surprisingly safe despite active climbing, swimming

58. (Age 3-5) Is there a Problem in that the Child Hits, Punches, Bites or Otherwise Injures Himself?

1. Yes, self only
2. Yes, others only
3. Yes, self and others
4. No (not a problem)

59. At What Age Did the Child Say His First Words (Even if Later Stopped Talking)?

1. Has never used words
2. 8-12 mo.
3. 13-15 mo.
4. 16-24 mo.
5. 2-3 years
6. 3-4 years
7. After 4 years
8. Don't know

60. (Before Age 5) Did the Child Start to Talk, Then Become Silent Again for a Week or More?

1. Yes, but later talked again
2. Yes, but never started again
3. No, continued to talk, or never began talking

61. (Before Age 5) Did the Child Start to Talk, Then Stop, and Begin to Whisper Instead, for a Week or More?

1. Yes, but later talked again
2. Yes, still only whispers
3. Now doesn't even whisper
4. No, continued to talk, or never began talking

62. (Age 1-5) How well Could the Child Pronounce His First Words When Learning to Speak, and How Well Could He Pronounce Difficult Words between 3 and 5?

1. Too little speech to tell, or other answer
2. Average or below average pronunciation of first words
3. Average or below on first words, usually good at 3-5
4. Unusually good on first words, average or below at 3-5
5. Unusually good on first words, and also at 3-5

63. (Age 3-5) Is the Child's Vocabular Greatly Out of Proportion to His Ability to Communicate?

1. Can point to many objects I name, but doesn't speak or communicate
2. Can correctly name many objects, but not communicate

3. Ability to communicate is pretty good
4 Doesn't use or understand words

64. When the Child Spoke His First Sentences, Did He Surprise You by Using Words He Had Not Used Individually Before?

1. Yes
2. No
3. Not sure
4. Too little speech to tell

65. How Did the Child Refer to Himself on First Learning to Talk?

1. "(John) fall down," or "Baby" (or Boy) fall down"
2. "Me fall down," or "I fell down"
3. "He, Him, She, or Her) fall down"
4. "You fall down"
5. Any combination of 1, 2, and/or 3
6. Combinations of 1 and 4
7. No speech or too little speech as yet

66. (Ages 3-5) Does the Child Repeat Phrases or Sentences That He Has Heard in the Past (Maybe Using a Hollow, Parrot-Like Voice), What is Said Having Little or No Relation to the Situation?

1. Yes, definitely, except voice not hollow or parrot-like
2. Yes, definitely, including peculiar voice tone
3. Not sure
4. No
5. Too little speech to tell

67. (Before Age 5) Can the Child Answer a Simple Question Like "What is Your First Name?" "Why Did Mommy Spank Billy?"

1. Yes, can answer such questions adequately
2. No, uses speech, but can't answer questions
3. Too little speech to tell

68. (Before Age 5) Can the Child Understand What You Say to Him, Judging from His Ability to Follow Instructions or Answer You?

1. Yes, understands very well
2. Yes, understands fairly well
3. Understands a little, if you repeat and repeat
4. Very little or no understanding

69. (Before Age 5) If the Child Talks, Do You Feel He Understands What He is Saying?

1. Doesn't talk enough to tell
2. No, he is just repeating what he's heard without understanding

70. (Before Age 5) Has the Child Used the Word "Yes"?

1. Has used "Yes" fairly often and correctly
2. Seldom has used "Yes" but has used "I"
3. Has used sentences, but hasn't used word "Yes."
4. Has used a number of other words/phrases, but not "Yes."
5. Has no speech, or too little speech to tell

71. (Ages 3-5) Does the Child Typically Say "Yes" by Repeating the Same Question He Has Been Asked?

1. Yes, definitely, does not say "yes" directly
2. No, would say "yes" or "OK" or similar answer
3. Not sure
4. Not enough speech to tell

72. (Before age 5) Has the Child Asked for Something by Using the Same Sentences You would Use When You Offer it to Him?

1. Yes, definitely (Uses "You" instead of "I")
2. No, would ask differently
3. Not sure
4. Not enough speech to tell

73. (Before Age 5) Has the Child Used the Word "I"?

1. Has used "I" fairly often and correctly
2. Seldom has used "I", but has used it correctly
3. Has used sentences, but hasn't used the word "I"
4. Has used a number of words or phrases, but not "I"
5. Has used "I", but only where word "you" belonged
6. Has not speech, or too little speech to tell

74. (Before Age 5) How does the Child Usually Say "NO" or Refuse Something?

1. He would just say "no"
2. He would ignore you
3. He would grunt and wave his arms
4. He would use some rigid meaningful phrase
5. He would use a phrase having only a private meaning
6. Other, or too li-tle to tell

75. (Before Age 5) Has the Child Used One Word or Idea as a Substitute for Another, for a Prolonged Time?

1. Yes, definitely
2. No
3. Not sure
4. Too little speech to tell

76A. Knowing What You Do Now, At What Age Do You Think You Might Have First Detected the Child's Abnormal Behavior?

1. In first 3 months
2. 4-6 months
3. 7-12 months
4. 13-24 months
5. 2-3 years
6. 3-4 years
7. After 4th year

76B. Knowing What You Do Now, At What Age Do You Think You Did First Detect the Child's Abnormal Behavior?

1. In first 3 months
2. 4-5 months
3. 7-12 months
4. 13-24 months
5. 2-3 years
6. 3-4 years
7. After 4th year

77. Father's Highest Educational Level

1. Did not graduate high school
2. High school graduate
3. Post high school technical training
4. Some college
5. College graduate
6. Some graduate work
7. Graduate degree

78. Mother's Highest Educational Level

1. Did not graduate from high school
2. High school graduate
3. Post high school technical training
4. Some college
5. College graduate
6. Some graduate work
7. Graduate degree

79. Number of Blood Relatives, Including Parents, Who Have Been in a Mental Hospital, or Who Were Known to Have Been Seriously Mentally Ill or Retarded, Diagnosed as Schizophrenia

    0. No relatives
    1. One relative
    2. Two relatives
    3. Three relatives
    4. Four relatives
    5. Five relatives

80. Number of Blood Relatives, Including Parents, Who Have Been in a Mental Hospital or Who Were Known to Have Been Seriously Mentally Ill or Retarded, Diagnosed as Depressive

    0. No relatives
    1. One relative
    2. Two relatives
    3. Three relatives
    4. Four relatives
    5. Five relatives

LIST OF REFERENCES

1. Ball, G., Classification Analysis, Technical Report, SRI 5533, Stanford Research Institute, Nov 1970.

2. Solomon, H., Numerical Taxonomy, Technical Report, ONR NR-042-267, Stanford University, Dec 1970.

3. Fortier, J. and H. Solomon, " Clustering Procedures," Multivariate Analysis, pp. 493-506, New York Academic Press, 1966.

4. Ball, G. and D. Hall, PROMENADE, An On-line Pattern Recognition System, Technical Report, TR-67-30, Stanford Research Institute, 1967.

5. Johnson, S., "Hierarchical Clustering Schemes," Psychometrica, v. 32, pp. 241-254, 1967.

6. Ward, J., "Hierarchical Grouping to Optimize an Objective Function," Journal of American Statistical Association, v. 58, pp. 236-244, 1963.

7. Fisher, R., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, v. 7, pp. 179-188, 1936.

8. Kanner, L., "Autistic Disturbances of Affective Contact," Nervous Child, v. 2, pp. 217-250, 1943.

9. Kanner, L., "Early Infantile Autism," Journal of Pediatrics, v. 25, pp. 211-217, 1944.

10. Rimland, B., Infantile Autism: The Syndrome and its Implications for a Neural Theory of Behavior, Appleton-Century-Crofts, 1964.

11. Bender, L., "Childhood Schizophrenia: A Clinical Study of One Hundred Schizophrenic Children," American Journal of Orthophychiatry, v. 17, pp. 40-56, 1947.

12. Bettelheim, B., "Schizophrenia as a Reaction to Extreme Situations," American Journal of Orthopsychiatry, v. 26, pp. 507-518, 1956.

13. O'Gorman, G., The Nature of Childhood Autism, Appleton-Century-Crofts, 1967.

14. Creak, M., and others, "Schizophrenia Syndrome in Childhood: Progress Report of a Working Party," British Medical Journal, v. 2, pp. 889-890.

15. Rimland, B., "The Differentiation of Childhood Psychoses: An Analysis of Checklist for 2,218 Psychotic Children," Journal of Autism and Childhood Schizophrenia, v. 1, 2, pp. 161-174, 1971.

16. Rimland, B., "High Dosage Levels of Certain Vitamins in the Treatment of Children with Severe Mental Disorders," To appear as a chapter in Orthomolecular Psychiatry, Freeman, 1971.

17. Wolfe, J., "Pattern Clustering by Multivariate Mixture Analysis," Multivariate Behavioral Research, v. 5, pp. 329-350, 1970.

INITIAL DISTRIBUTION LIST

	No. Copies
~~1. Defense Documentation Center Cameron Station Alexandria, Virginia 22314~~	~~2~~
2. Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3. Professor J. R. Borsting, Code 55 Bg Chairman, Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940	x 3
~~4. Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940~~	~~1~~
~~5. Chief of Naval Personnel Pers 11b Department of the Navy Washington, D. C. 20370~~	~~1~~
~~6. Dr. Bernard Rimland Institute for Child Behavior Research 4758 Edgeware Road San Diego, California 92116~~	~~1~~
7. LCDR Earnest W. Richardson, USN 1210 Abernathy Street Fordyce, Arkansas 71742	1

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author) Naval Postgraduate School Monterey, California 93940	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP
REPORT TITLE A Hierarchical Clustering Technique	
DESCRIPTIVE NOTES (Type of report and, inclusive dates) Master's Thesis; September 1971	
AUTHOR(S) (First name, middle initial, last name) Earnest Wells Richardson	
REPORT DATE September 1971	7a. TOTAL NO. OF PAGES 7b. NO. OF REFS 17
a. CONTRACT OR GRANT NO. b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
DISTRIBUTION STATEMENT Approved for public release; distribution unlimited.	
1. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Naval Postgraduate School Monterey, California 93940

1. ABSTRACT

This thesis summarizes statistical clustering procedures and presents in some detail a hierarchical clustering technique and computer routine which utilizes Euclidean distances as measures of object similarity. An application of the technique is made to scores derived from a qualitative data base describing mentally disturbed children, and results of the application are compared to results obtained from previous clustering studies.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED

Security Classification

A-31409

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Clustering						
Pattern recognition						
Classification						
Qualitative data						
Factor analysis						
Homogeneous groups						
Similarity measures						
Psychotic child						
Childhood autism						
Vitamin treatments						